EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

ICTPSP
ICT POLICY SUPPORT PROGRAMME
part of the Competitiveness and Innovation Framework Programme CIP

DEMO

Project Acronym: EuDML
Grant Agreement number: 250503
Project Title: The European Digital Mathematics Library

D8.3: Toolset for Entity and Semantic Associations – Value Release

Revision: 1.0 as of 31st May 2012

Authors:

Mark Lee	University of Birmingham, UB
Petr Sojka	Masaryk University, MU
Radim Řehůřek	Masaryk University, MU

Contributors:

Josef Baker	University of Birmingham, UB

Project co-funded by the European Comission within the ICT Policy Support Programme

Dissemination Level		
P	Public	✓
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.1	12th March 2012	Petr Sojka	MU	First version as placeholder for partner's input.
0.2	22nd May 2012	Mark Lee	UB	First attempt at draft.
0.21	22nd May 2012	Josef Baker	UB	Minor edits.
0.3	25th May 2012	Petr Sojka	MU	Gensim, multilinguality.
0.4	28th May 2012	Mark Lee	UB	Citation Indexing, Summary.
0.5	28th May 2012	Petr Sojka	MU	Gensim implementation.
0.6	30th May 2012	Mark Lee	UB	Minor Edits.
0.7	30th May 2012	Petr Sojka	MU	Final overview.
1.0	31st May 2012	Petr Sojka	MU	Final release candidate.

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

Executive Summary
In this document we describe the value release of the toolset for entity and semantic associations, integrating Unsupervised Document Similarity implemented by MU (using gensim tool) and Citation Indexing and Matching (as provided by ICM and UJF/CMD). We give a brief description of tools and provide some initial evaluation.

1 Introduction

In D8.2 [7] we released a prototype toolkit of services which facilitated the discovery of semantic associations between documents within our collection. The toolkit focused on two key techniques:

Similarity Clustering tries to collate articles in collection with similar topic and content. It can be achieved either by automatically classifying documents according to a pre-defined scheme (e.g. MSC) or clustering documents based on co-occurring terms.

Citation Indexing aims to create a network of documents within the collection by automatic parsing and linking of citations.

The toolkit consists of two services for each functionality. These are:

1. Semantic Similarity
   - gensim (MU) [9] is a software library implementing e.g. unsupervised clustering algorithm which uses various machine learning methods to measure semantic similarity based on word co-occurences (distributional semantics).
   - Yadda Similarity Service (ICM) a service based on the Lucene open source information retrieval software library [8].
2. Citation Interlinking
   - Yadda Citation Interlinking Service (based on the Lucene open source information retrieval software library [8]).
   - UJF Citation Matcher (a citation resolution service based on [4]).

This document describes our value release of the association toolkit and some initial evaluation. In Section 2 we describe document classification and clustering and in Section 3 we describe tools for citation indexing and matching.

2 Similarity Services

Similarity service delivers functionality of finding ordered set of documents semantically similar to given one. Exact features of documents which are considered during semantic similarity search depend on implementation—we have implemented several state-of-the-art/leading-edge methods.

The service basically has two operations:

- **Similarity indexing –** All documents must be ‘indexed’ before being made available to similarity search. Indexing is done only once for each document and so it can be considered as a setup process for the service.

Aspects of the indexing which may vary between implementations is the ability to add documents in incremental fashion: incremental indexing means that document collections may be added to the service without the need to reindex all the documents added before. Availability of incremental indexing (on-line processing) is one of features which should be considered during the comparison of different implementations.

- **Similarity search** The core functionality of semantic similarity search is to find similar documents to particular document. Given a set of documents which are indexed (i.e. the collection), the similarity service should return an ordered list of documents which are similar to the specified one. It should be noted that different similarity metrics may be defined (in fact each implementation of the service is a realization of a particular metric) and so a comparison of different implementations will heavily depend upon the chosen similarity metrics.

After giving a general overview of the service we focus on two implementations in the toolkit: GENSIM library and Yadda similarity service implementation.

2.1 GENSIM

The goal of this tool is to assist humans in data exploration via similarity browsing. GENSIM [3] is a software framework for modelling semantic similarity of text documents. Within the context of digital libraries, it allows querying for “similar documents”, with similarity based purely on document contents (i.e. plain text, or abstract with metadata).

2.1.1 Gensim Demos

To demonstrate the capabilities of GENSIM, an Internet-based demos have been produced which showcases possible scenarios (similarity demos):

Proof-of-concept Demo 1 of GENSIM possibilities. Four subdemos available at `http://aura.fi.muni.cz:8080` demonstrate GENSIM functions to answer questions:

1. For a given article, what are its ten most similar articles in the library?
2. Given two articles, how similar are they?
3. What are the pairs of the most similar articles across the entire collection (plagiarism candidates)?
4. Given an article, what are the topics/‘keywords’ covered by this article (data exploration)?

MIaS4gensim Demo 2 `http://aura.fi.muni.cz:8889` This Demo showcases most and least common document terms and most prominent words for LDA and LSA topics. From the results it is clear that math formulae subterms frequently appear as the topics and thus are important to consider in math-aware document similarity computations like EuDML ones. As a result, math should be taken into account in building paper’s bag-of-word list during tokenization for EuDML type of similarity handling to bring math-awareness.

EuDML integration Demo 3 shows integration of GENSIM in beta version of EuDML system. The deployed GENSIM similarity system can be found at `http://eudml.org`, where it is used to compute similarities for more than 100,000 documents, on-the-fly.

As not enough EuDML data was available at the time of the demo preparation and testing, first two demo URLs use data from the MREC collection of mathematical texts (a snapshot of 434,894 full-text articles from ARXMLIV, originally from ARXIV) to verify the framework's usage and scalability on documents typical for EuDML (scientific papers with lot of math).

2.1.2 Statistical Semantics

GENSIM contains several automated algorithms for deriving semantic representation from plain text, and therefore a choice is given to demo users between Latent Dirichlet Allocation (LDA) [1], Latent Semantic Analysis (LSA) [2] and plain TF-IDF in subemos 1–3 of GENSIM Demo 1, by means of a drop-down list. The first two methods compute a higher-level, semantic similarity, the last one only measures (weighted) word overlap.

Semantic properties of LSA and LDA come from exploiting word co-occurrence within documents. In a training corpus of documents, words that tend to appear together are taken to be semantically related and soft-clustered together ("soft" because a word belongs to each cluster with a weight, or probability, not as a binary decision). Each such cluster of words describes one topic. Obtaining the clustering automatically and efficiently is a major challenge; GENSIM is a leading framework in scalable model training.

Given a trained LSA or LDA model, any text document can be described by how much it belongs to each topic. This gives a higher level (more abstract) representation of the document's contents—now even two documents that do not share any words in common can be evaluated as closely similar. This is a strict departure from an exact keyword overlap as popularized in search engines with boolean keyword searches.

2.1.3 Deployment of Gensim at EuDML

To integrate GENSIM (in Python) into Java-based EuDML several steps were undertaken. GENSIM runs as client/server architecture. Server stores similarity index and client does similarity search itself (computes results on-the-fly for a given document).

For adding new documents there were implemented transaction handling (similarity index is copied at the beginning of transaction). Similarity index is created for every language separately. Language having more than 1000 documents uses TF-IDF (log entropy) model, for languages with more documents LSA with 400 dimensions (topics) is currently used. Standard tokenization is done (only alphabetical terms from both full-text and metadata are currently used).

2.2 Yadda Similarity Service

The Yadda similarity service is implemented over the Lucene full-text search engine [8] and its "more like this" search functionality. Indexing consists of building an inverted index of documents, i.e. a mapping between words (terms) and the documents which contain them together with some additional statistics (for instance the number of occurrences in each document). The inverted index is used for similarity searches in the following fashion:

- A given document's important features are determined during search. For full-text search, any 'feature' is just a word which can be considered specific or highly significant for the given document. Such words are chosen using term frequency/inverse document frequency (TF-IDF) statistics for each word.

- The most specific/significant words of the document are used to construct a boolean OR full-text query—standard full-text search and its "term vector model" is used to determine set of documents which match the query in the best way (documents which are most similar to the words in the query). Documents found during full-text search are returned as similarity results.

2.3 Comparing Gensim with Yadda Similarity Service

An initial comparison of the two services was conducted by randomly selecting six documents (two English, two German and two French articles) and comparing the top five "most similar" articles returned by GENSIM and the Yadda similarity service.

For example, "Minoru Itoh (2000) Capelli elements for the orthogonal Lie algebras. Volume 10: Issue 2, Publisher page 463–489" [5] (eudml:121658) results in the following articles being returned by GENSIM (ordered by similarity):

- On the nilpotency of certain subalgebras of Kac-Moody Lie algebras. Kim, Yeonok, Misra, Kailash C., Stitzinger, Ernie
- A Leibniz algebra structure on the second tensor power. Kurdiani, R., Pirashvili, T.
- Lie quasi-bialgebras with quasi-triangular decomposition. Andruskiewitsch, Nicolás, Tiraboschi, Alejandro
- Surprising properties of centralisers in classical Lie algebras Oksana Yakimova
- A real analog of Kostant's version of the Bott-Borel-Weil theorem. Šilhan, Josef

Whereas the Yadda similarity service produces the following articles:

- On commutation relations for 3 3-graded Lie algebras. de Oliveira, M.P.
- Stem extensions and stem covers of Leibniz algebras. Casas, J. M., Ladra, M.
- A real analog of Kostant's version of the Bott-Borel-Weil theorem. Šilhan, Josef
- A Leibniz algebra structure on the second tensor power. Kurdiani, R., Pirashvili, T.
- Surprising properties of centralisers in classical Lie algebras Oksana Yakimova

As expected there is an overlap between the sets of documents returned. However, in general we have found GENSIM to return more subject specific research articles rather than articles consisting of a generic overview of a wider subject. In addition, GENSIM appears to cope better with different languages. Disappointingly, one of the two French language articles and both of the German language articles produced no similar articles according the Yadda similarity service. As described in D8.2 [7], the Yadda similarity service maintains language specific indexes and this might be the source of the problem but our analysis is ongoing.

As a result, only GENSIM has been deployed in the current EuDML system (version 1.3). It is sufficiently robust even though for some articles there are no full-texts available, and thus GENSIM is using only basic metadata and MSC tokens for bag-of-words semantic paper representation. It remains to be seen whether it gives usable results even in these cases, or whether similarities will not be exposed for papers without full-texts at EuDML disposal.

The issue of the export of similarity results via service (either for data providers, or as Linked data) is worth to consider. List of similar papers might possibly be exposed as OpenSearch result ordered sets.

3 Linking and Matching Tools

Drawing links between articles in a collection is important to support the rapid detection and retrieval of articles similar to a given input article or query and to serve the results to the user. Since many documents in the EuDML collection are likely to contain bibliographic references, and many of these references are likely to point to (other) documents within the EuDML collection we are aiming to use *Bibliographic Reference Matching* as the primary means to build up a relationship network between articles. The goal of bibliographic reference matching is to assign to a bibliographic reference an identifier of the referenced document.

3.1 Yadda Citation Interlinking Service

Bibliographic reference matching is implemented as a set of *processing nodes*. One node creates document metadata for each bibliographic reference. Another one matches a given document and its bibliographic references with relevant entities in the index. Yet another node updates NLM entries with matched document identifiers.

Matching of a document or bibliographic reference is performed by a series of queries to the metadata index. Firstly, if we know an identifier of the entity being matched, such as DOI, MR or Zbl, we query the index for documents with one of these identifiers. Otherwise, we query the index for documents given authors' surnames, a hash function of journal title, and year. For each of the hit, we check the remaining fields (possibly, as in the case of titles, using string distance functions). If we still cannot find matching entities, we weaken the query conditions to surnames and year alone, repeating the evaluation of hits.

Bibliographic reference matching is implemented by two chains of processing nodes. In the first process, the source node `DateRangeItemRecordIteratorBuilder` iterates over all the item records in the repository. The next node, `ItemRecordToYElementConverterNode` accepts such item records on input and converts them to format accepted by indexing module. Next, `RelationsToElementsExpanderNode` takes bibliographic references in the input metadata records and promotes them to "first-class" records. This way all the bibliographic references can be indexed in the same way in which the main records are indexed in the system.

The second process also iterates over the item records in the repository. Next, using `ItemRecordToEnhancerMessageNode`, it wraps the records in messages that are used in further processing. Finally, the messages are piped to `BibReferenceMatchingWriterNode`, which matches the given document with other documents indexed in the system, and stores the results of the matching in the NLM format in `ext-link` element with attribute `ext-link-type` set to `eudml-item-id`.

3.2 UJF Citation Matcher

As described in D8.1 [6], D8.2 [7] and [4], the UJF Citation Matcher provides a robust method for resolving (incomplete or possibly incorrect) citations to a particular document or identifier.

The UJF/CMD citation extraction and matching algorithm does not attempt to perform citation parsing or citation field tagging prior to trying to find matching citations. Instead, citations are viewed simply as strings of characters with no attempt to parse a structure to the citation string. As argued in [4] this avoids several problems:

1. Even if citation parsing is successful, individual fields often remain coded in different ways and cannot be compared using exact comparison methods.
2. Errors in parsing can result in the complete failure of the matching process.
3. The parsing process is both costly and error prone.

For these reasons, the UJF/CMD approach relies on just a shallow analysis of the input string and relies on a number of string similarity evaluation methods and ad hoc heuristics which work reasonably well in practice in the context of mathematical databases. In particular the matcher relies on numerical information in the citation string to resolve the citation. It is expected that such a method is relatively resistant to multilingual and typesetting issues.

As reported in D8.2, as part of the toolset, we offer three online demos of the UJF Citation Matcher:

- At `http://thar.ujf-grenoble.fr/cgi-bin/eulookup`, an interactive lookup where the user inputs a bibliographic citation (as a string) and gets back near matches when they are found.
- At `http://thar.ujf-grenoble.fr/cgi-bin/batcheulookup` a demo of a batch tool using the same matching engine.
- At `http://atlas.ujf-grenoble.fr/cgi-bin/zlookup` an example implementation of the reference matcher for the Zentralblatt Math database can be found.

3.3 Comparing Citation Matching and Interlinking Services

Detailed comparison and evaluation of the two approaches to citation interlinking and matching are ongoing. However we can report some initial findings. Both systems were run on a sub-collection of the XML metadata for 190,000 documents to attempt to resolve any citation within the metadata to a pre-existing item within the EuDML collection. The results were as follows:

Yadda Citation Interlinking Service: 16,996 citations
UJF Citation Matcher: 58,534 citations
Classified by both: 15,708 citations.

In other words, the UJF Citation Matcher appears to have a far higher recall which can be explained by the more robust method of citation matching adopted. However, there is a small but significant number of citations which are not resolved by this method but which are by the Yadda service. Of course these results require further analysis and in particular, a measure of precision or accuracy of the resolved citations is required. This will be done as part of the planned internal evaluation of EuDML Task 8.3.

4 Evaluation and Further Work

Since a thorough evaluation of the services is planned in task 8.3 of workpackage 8 we will limit ourselves to only some basic observations in this section. They primarily serve to point out some of the further work that will need to be done on the integrated services.

Multilinguality Problems While the content provided in EuDML includes articles in a large number of languages there is not necessarily a large number of articles for every language. However, most of the techniques we integrate in this workpackage depend on a sufficiently large sample set to work with.

Consider, for instance, the GENSIM tool, which primarily works with demo data obtained from the ARXIV. ARXIV contains a majority of English articles but considerably less in other languages. For example, when looking for documents that are similar to an article in French one can easily obtain a number of very close matches since all French articles are attributed to a single topic and the only difference can stem from the respective English abstracts.

As multilinguality is a key feature in EuDML these issues will have to be addressed. Possible ideas to try:

Math formulae as interlingual paper representation As described in D7.3 [11], we make every effort to get mathematical formulae from both born-digital (by maxtract) or scanned (by Infty) papers. Formulae can be viewed as language independent representation (interligual) of paper. We have done experiments with gensim using weighted tokens generated by MIaS [10] for math formulae Lucene indexing as the 'weighted math bag of word' representation. This could be used as interlingual part of paper index, in addition to paper MSC codes, author names and possibly (low weighted) official journal name.

interlingual keyword list As part of the gensim demo there is a subdemo number 4 at `http://aura.fi.muni.cz:8080/` which shows gensim's capability to offer keywords for a given paper. Keywords can be computed for paper, collection of papers (e.g. sharing the same MSC, or authored by the same author). The amount of keywords and collocation detection could be set up by preprocessing and finetuning the gensim parameters. With this approach, one can get keyword/collocation lists for given author of MSC. These lists might be Google translated to some 'interlingua' (English, or French :-), and checked by humans. This interlingual word list for given MSC and/or author and/or paper could be added to the interlingual bag of word of paper for similarity computations.

As manual checking of translations is time-consuming expert task, we fear that this is out of reach of current EuDML project.

interlingual automated translation Google-translated full-text might be used as solution of multilingual information retrieval task.

Similarity Enhancements Currently, there are publications without full-texts in EuDML. Even though methods used by gensim library are robust to noise, comparing papers with basic metadata to those with full-texts is particularly challenging. Even though paper length normalization does implicitly take place (and increasing rôle of words in authors, title and MSC code fields), results may not be optimal. GENSIM parameters might be fine-tuned and explicit weighting may be used to resolve the discrepancy. Also comparing only among the sets of full-text papers and sets of papers without full-texts may be acceptable solution.

Bibliographic Reference Metrics play an important rôle in the matching tools we provide. Therefore, thorough experimentation and evaluation of different metrics as well as combinations of metrics will have to be carried out. In the current implementation the used metrics are relatively strict and often far too syntax driven to allow for proper semantic association.

For example, when experimenting with bibliographic reference matching on the Zentralblatt database, one can easily observe some curious artifacts. Occasionally close matches for a particular article are ranked higher than the actual article. Similarly, the distance metric is particularly fragile with respect to changes in journal names. It gives higher rankings if the journal is given in the correct abbreviation while giving (sometimes significantly) rankings if the journal names are given in full or differently or only partially abbreviated.

This already indicates that both metrics and selected training data is currently too restrictive and has to be broadened in order to achieve semantically more appropriate results.

Query Corrections Currently the integrate tools will primarily be employed to produce enhanced metadata and pre-compute association networks in order to serve the users with articles related to a particular query or input document. However, in the light of EuDML's aim to improve accessibility one could imagine that similarity matching could also be exploited to provide query correction, e.g. for misspelled mathematical expressions, to support print impaired users. The similarity techniques could enable us to base these corrections on a mathematical corpus rather than using a standard dictionary and would require the introduction of some further degree of fuzziness into the matching algorithms. Clearly this part of the workpackage will then overlap with WP10.

5 Summary, Conclusions

This document reports the value release of the toolset for entity and semantic associations, integrating Unsupervised Document Similarity implemented by MU (using GENSIM tool) and Citation Indexing and Matching (as provided by ICM and UJF/CMD). The toolset consists of stable technology which is both integrated within EuDML and stand-alone demos. The next step is to provide a more indepth evaluation of the technologies in Task 8.3.

References

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[3] gensim. `http://nlp.fi.muni.cz/projekty/gensim/`.

[4] Claude Goutorbe. Document Interlinking in a Digital Math Library. In Petr Sojka, editor, *Towards a Digital Mathematics Library*, pages 85–94, Grand Bend, Ontario, Canada, 2008. Masaryk University, Brno.

[5] Minoru Itoh. Capelli elements for the orthogonal Lie algebras. *J. Lie Theory*, 10(2):463–489, 2000.

[6] Mark Lee, Petr Sojka, Volker Sorge, Josef Baker, Wojtek Hury, and Łukasz Bolikowski. Association Analyzer Implementation: State of the Art, November 2010. Deliverable D8.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, `https://www.eudml.eu/sites/default/files/D8.1_0.pdf`.

[7] Mark Lee, Petr Sojka, Volker Sorge, Wojtek Hury, Łukasz Bolikowski, and Radim Řehůřek. Toolset for Entity and Semantic Associations – Initial Release, May 2011. Deliverable D8.2 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, `https://www.eudml.eu/sites/default/files/D8.2_0.pdf`.

[8] Apache Lucene. `http://lucene.apache.org/`.

[9] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`, software available at `http://nlp.fi.muni.cz/projekty/gensim`.

[10] Petr Sojka and Martin Líška. Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In James H. Davenport, William M. Farmer, Josef Urban, and Florian Rabe, editors, *Intelligent Computer Mathematics. Proceedings of 18th Symposium, Calculemus 2011, and 10th International Conference, MKM 2011*, volume 6824 of *Lecture Notes in Artificial Intelligence, LNAI*, pages 228–243, Berlin, Germany, July 2011. Springer-Verlag. `http://dx.doi.org/10.1007/978-3-642-22673-1_16`.

[11] Petr Sojka, Krzysztof Wojciechowski, Nicolas Houillon, Michal Růžička, and Radim Hatlapatka. Toolset for Image and Text Processing and Metadata Enhancements – Value Release, March 2012. Deliverable D7.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, `https://www.eudml.eu/sites/default/files/D7.3.pdf`.